Unsupervised Active Learning For Video Annotation

Emre Demir EMREDEMIR@ITU.EDU.TR
Department of Computer Science, Istanbul Technical University, Istanbul, Turkey

Zehra Cataltepe CATALTEPE@ITU.EDU.TR
Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

Umit Ekmekci UEKMEKCI@ITU.EDU.TR
Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

Mateusz Budnik MATEUSZ.BUDNIK@IMAG.FR
University of Grenoble Alpes, LIG, F-38000 Grenoble, France

Laurent Besacier LAURENT.BESACIER@IMAG.FR
University of Grenoble Alpes, LIG, F-38000 Grenoble, France

Abstract

When annotating complex multimedia data like videos, a human expert usually annotates them manually. However, labeling these immense quantities of videos manually is a labor-intensive and time-consuming process. Therefore, computational methods, such as active learning are used to help annotate. In this study, we propose a cluster based unsupervised active learning approach and a new active learning method for unsupervised active learning on REPERE (Giraudel et al., 2012) video dataset, which is created for the problem of person identification in videos. Our study aims to identify who is speaking and who is on screen by using multi-modal data.

1. Introduction

Annotating immense quantity of videos manually is a labor-intensive and time-consuming process. On the other hand, automatic annotation techniques such as Active Learning offer various solutions to overcome the excessive cost of manual annotation. Different active learning methods have been proposed for the video annotation problem. One of the studies (Ayache & Quénot, 2007) applies active learning for video annotation by comparing uncertainty sampling, the most probable sampling and random sampling for video indexing. (Ayache & Quénot, 2008) proposes video retrieval and annotation system called LIGVID which uses two active learning methods: 'relevance sampling' and 'uncertainty sampling'. Another study integrates SVM based active learning for feature selection to solve the text classification problem (Joshi et al., 2006). However, active learning for feature selection fails in that study, because of the use of a wrong feature reduction technique called GainRatio Feature Selection (Joshi et al., 2006).

In the study (Bilgic et al., 2010), active learning is applied on a networked data, nodes of which are 'papers' and links are 'references to the other papers'. It uses a method based on query by disagreement and reduces paper annotation costs for classifying research papers. The study (Raghavan et al., 2006) extends the traditional active learning framework by including feedback on features alongside labeling the instances. It focuses on the effects of feature selection and human feedbacks for features in the setting of text categorization and applies uncertainty sampling based methods.

In this study[1], we propose a cluster based unsupervised active learning approach as a selection strategy on REPERE (Giraudel et al., 2012) video dataset, which is created for the problem of person identification in videos. Our study aims to identify who is speaking and who is on screen.

[1]The study is a part of the project CAMOMILE (cam), which targets to produce an annotation framework for multimodal, multimedia and multilingual data

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

2. The Properties of The Dataset

We use the video dataset from the REPERE challenge, which aims to find answers to questions "Who is speaking?", "Who is present in the video?", etc., by the use of various information on speech and image extracted from the dataset. The REPERE Corpus consists of 28 videos, which includes 7 different types of shows such as news, talk show, etc. and various numbers of participants from 3 to dozens in a video. Furthermore, the length of the videos has a range of 3 to 30 minutes, which naturally causes various numbers of annotations for each video approximately 20 to 100 frames.

(Budnik et al., 2014) applies speech and face segmentation processes on videos to gather the similarity matrices face-to-face and speech-to-speech which are normalized into the interval [0,1]. The third similarity matrix, face-to-speech, occurs from correlation scores between the faces and the speakers to build a multimodal clustering. We use these three similarity matrices for training and manually annotated videos from REPERE dataset for testing. The study (Poignant et al., 2012) extracts overlaid texts in videos by an Optical Character Recognition (OCR) system to gather an initial set of annotated data. We use the information from embedded texts in videos that can point the name of a represented speaker.

3. Notation

In this study we use the following notation. The finite data set is $X = x_1, x_2, \ldots, x_n$ where the cardinality is $|X| = n$. The cluster set $C = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_k\}$ represents the cluster sets of the set X with the assumption $|\mathrm{C}_i| > 0$ for all $i = 1, 2, \ldots, k$. C' is the second clustering for X where $C' = \{\mathrm{C'}_1, \mathrm{C'}_2, \ldots, \mathrm{C'}_\ell\} \in S(X)$. $\mathrm{M} = m_{ij}$ is the confusion matrix of each clustering pair C, C'. The intersection between C_i and $\mathrm{C'}_j$ is a $kx\ell$ matrix where the ij^{th} element gives the number of the elements in the intersection of C_i and $\mathrm{C'}_j$.

$$m_{ij} = |\mathrm{C}_i \cap \mathrm{C'}_j|, 1 \leq i \leq k, 1 \leq j \leq \ell \tag{1}$$

4. Unsupervised Active Learning

Supervised query by disagreement method (QBD) (Settles, 2012) is one of the active learning methods in the literature. QBD uses two different classifiers and asks the label of a disagreed point to an expert. We introduce unsupervised query by disagreement method to accelerate the learning phase of the automatic video annotation by using unsupervised learners. When it is used with supervised learners, Query by Disagreement (QBD) lets the learners label each instance and compares the learners' outputs for the same instance to detect disagreement instances. However, since unsupervised methods do not generate labels for each data instance as output of a learner, using QDB on unsupervised learners is a new and a challenging problem. We propose a novel approach for QBD on unsupervised learners and apply the solution on multimodal video data.

In our study, we use the following steps during an Active Learning phase: clustering, cluster matching, disagreement measurement between clusters, selection of the most disagreed clusters, selection of an instance to be queried from the selected cluster.

4.1. Clustering and Cluster Matching

Active learning cycle begins with the clustering of the data. We use two clustering algorithms: Agglomerative Clustering and K-Medoid Clustering. For a given threshold, Agglomerative Clustering estimates the number of clusters, which we use as the value of the K-Medoid Clustering's number of clusters parameter k.

In order to find the clustering disaggreement, we first solve the complementary problem of finding the disagreements between clusterings. We apply a cluster matching algorithm to measure the similarities between the clusterings produced by the Agglomerative and K-Medoid algorithms. We use the intersection of clusters by calculating the cluster similarity metric as in equation 2.

$$H(C, C') = \frac{1}{k} \sum_{j=match(i)} m_{ij} \tag{2}$$

We need to find $j = match(i)$ which denotes the cluster that the cluster C'_j to which the C_i should be matched. Then, in order to assign a clusters between Agglomerative and K-Medoid clusterings, we adopt and apply the Gale-Shapley algorithm (Iwama & Miyazaki, 2008; Krumpelman & Ghosh, 2007) which is used for solving the 'Stable Marriage Problem'. Gale-Shapley algorithm guarantees that the solution obtained is perfect (i.e. everyone gets married) and stable.

Let the clustering produced by the Agglomerative clustering be $C = C_1, C_2, \ldots, C_k$ and the clustering produced by the k-medoid clustering be $C' = C'_1, C'_2, \ldots, C'_k$. Gale-Shapley algorithm requires that each cluster C_i ranks the clusters C'_j and vice versa. We use the cluster confusion matrix entries (Equation 1) in order to produce these rankings and then apply the Gale-Shapley algorithm to produce a cluster matching. The output of the Gale-Shapley is a matching $M = [m_1, m_2, \ldots, m_k]$ where, $m_i \in 1, \ldots, k$, $m_i = j'$ if C_i is matched with C'_j in the stable matching.

4.2. Query Instance Selection

In order be able to select which instance to query, we first select the most informative cluster of cluster pair, and then we select the most informative instance in the selected clus-

ter.

4.2.1. Cluster Selection

We score clusters using different methods and select and instance from the cluster with the highest score. ***Big Cluster Selection (BCS):*** The study (Budnik et al., 2014) proposes 'Big Cluster First' selection strategy which calculates a score using size of a set and the number of annotated instances in that set. The method selects an instance from a minimum scored set by asking human expert. The BCS strategy score for a cluster C_i is calculated as $BCS(C_i) = NumberOfAnnotations(C_i)/Size(C_i)$.

The Most Disagreement Selection (MDS):In theory, the most disagreed cluster pair gives us the most uncertain points because two stable matched clusters have lots of disagreed instances. For measuring the disagreement between a pair of matched clusters C_i, C'_j where $j = m_i$, we divide the number of common instances by the total number of instances:

$$MDS(C_i) = \frac{C_i \cap C'_j}{C_i \cup C'_j} \tag{3}$$

MDS method choosed the cluster C_i which has the highest DS score.

Hybrid Cluster Selection BCS method performs better during the initial stages of active learning where the number of labeled instances are very few and labeling instances on big clusters help a label a lot of instances. On the other hand, MDS performs better when clusters contain more known instances. Because cluster disagreement has a correlation with the label assignment. Therefore, we introduce two hybrid cluster selection methods: Soft Hybrid Selection (SH) and Hard Hybrid Selection (HH).

In order to combine or compare the different BCS and MDS clustering scores, we apply the z-score normalization on them.

The Soft Hybrid Selection (SH) score for a cluster C_i is computed as the weighted average of the normalized BCS and MDS scores:

$$SH(C_i) = ((1 - \alpha) \times BCS(C_i)) + \alpha \times MDS(C_i) \tag{4}$$

where $0 < \alpha < 1$. Since we want to give more weight to the BCS method during the initial iterations and more weight to MDS during the later iterations, we vary the value of the weight $alpha$ by passing through the $sigmoid$ function in each iteration number t.

The Hard Hybrid Selection method uses the BCS scores at the beginning of active learning iterations and then uses the BCS scores after a certain iteration number five.

4.2.2. Instance Selecting Strategy

Instance selection strategy tries to determine the most informative instances in the selected cluster. The instances from the center of a cluster are more 'certain' than the instances close to the cluster boundary in terms of class knowledge. However, since the entropies of uncertain instances are relatively higher than the entropies of certain instances (Settles, 2012), the most informative instances are actually on the region around the boundary of a cluster. Therefore, for each instance in a cluster, we sum the distances to other instances in the same cluster and choose the instance with the highest sum, which gives us the most 'uncertain' instance, the instance which is farthest away from all the other instances. Instance selection strategy can select the most 'uncertain' or the most 'certain' (i.e. medoid) instance from a cluster or a cluster pair.

5. Results and Discussion

We evaluate the performance of selection strategies over active learning cycles using multimodality on 28 videos from 7 different TV programs. Each video has three different similarity matrices namely face-to-face, speech-to-speech and face-to-speech. We run four experiments with regard to matrices in this order; 'face score for face track annotation" (FF), 'face score for speaker track annotation' (FS), ' speaker score for speaker track annotation' (SA) and 'speaker score for speaker annotation' (SS). An active learning cycle which is depicted as 'step', asks one annotation for each video. As the performance measure, we use the F-measure, instead of accuracy, since the number of instances of each class (person) in the datasets is very different from each other. F-measure is the harmonic mean of precision p and recall r values. Precision shows how much relevant instances are retrieved among all retrieved instances and recall shows how much retrieved instances are relevant among all relevant instances.

Figure 1 (a) shows the F-measure values on the FF in each step. The results indicate that active learning methods BCS -uses uncertainty- and Hard Hybrid perform better than the random method in the earlier steps. However, random selection achieves the best values among all other methods at the later steps. On the other hand, MDS-Uncertainty, MDS-Certainty and Soft Hybrid achieve a little better performance than random for SS as shown in Figure 1 (d). Soft Hybrid selection performs worse in FF, FS and SF and barely good at SS.

The multimodal classifying problems FS and SF (See Figure 1 (b),(c)) have more deviations than the others because the correlation between face tracks to speaker tracks is weaker than head to head tracks or speaker to speaker tracks. For this reason, MDS certainty performs better than all others in FS and more robust than BCS in SF.

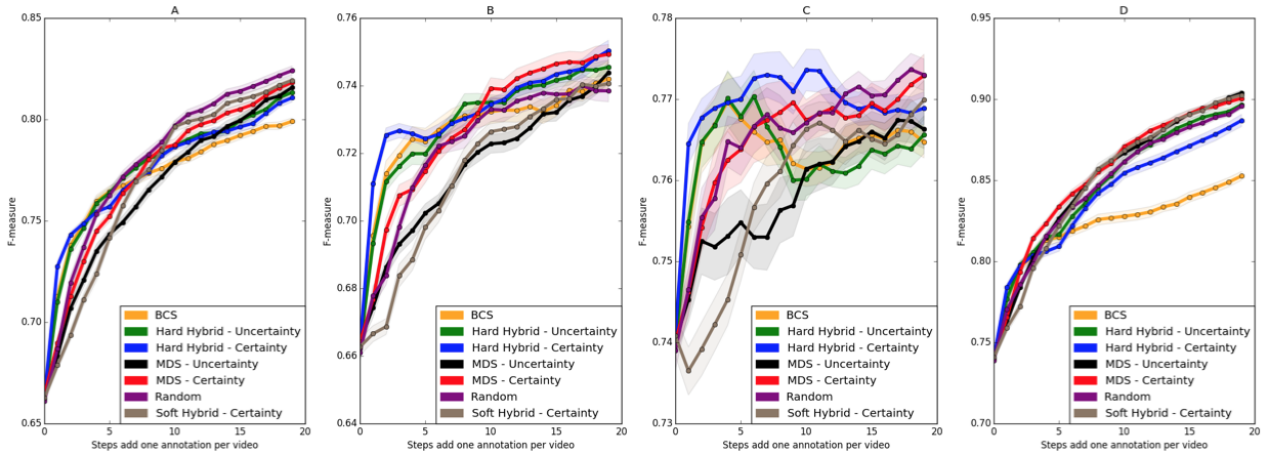

Figure 1. F-measure scores, from left to right: (a)face score for face modality (FF), (b)face score for speaker modality (FS), (c)speaker score for face modality (SF) and (d)speaker score for speaker modality (SS).

In the experiment FS, BCS selects good annotations at the earlier steps and achieves high F-measure value more rapidly than MDS-Certainty since there is no sufficient data to be disagreed upon for the MDS-Certainty method. However, at the later steps, F-measure value decreases, but MDS-Certainty keeps increasing. In order to take advantage of the strongest sides of both methods, Hard Hybrid Certainty (HHC) use BCS at the first five steps. After the fifth step, it uses MDS and keeps increasing in terms of F-measure. As a result, HHC achieves better scores than random.

In the experiment SF, BCS achieves high F-measure scores rapidly and higher than random as in FS. Nevertheless, it's decreasing at the later steps and finally performs worse than random. On the other hand, MDS-Certainty F-measure increases stably as in FS but performs similar to random. However, the most interesting result comes from HHC that combines BCS and MDS-Certainty. HHC gives higher F-measures rapidly like BCS at the earlier steps and continues to increase at the later stages like MDS-Certainty. Furthermore, it performs better than random significantly until the 10th step. Fortunately, in terms of active learning, reaching to the highest F-measure score rapidly is more meaningful.

We proposed the MDS active learning method and its hybrid variations and we applied them on multimodal video annotation data. According to our experiments, for different types of annotation tasks, different active learning strategies could be more suitable. Hybrid strategies could be more successful than using a single strategy alone. Decision of the cluster selection and instance selection method adaptively during each active learning step, using a synthetic dataset to investigate the merits of these strategies, examination of each method for each video, rather than the whole REPERE corpus are the future research directions we aim to follow.

Acknowledgments

Work in this paper is partially supported by the Scientific and Technological Research Council of Turkey (TUBITAK) project 112E176.

References

Collaborative annotation of multi-modal, multi-lingual and multi-media documents. URL http://www.chistera.eu/projects/camomile.

Ayache, Stéphane and Quénot, Georges. Evaluation of active learning strategies for video indexing. *Signal Processing: Image Communication*, 22(7):692–704, 2007.

Ayache, Stéphane and Quénot, Georges. Video corpus annotation using active learning. In *Advances in Information Retrieval*, pp. 187–198. Springer, 2008.

Bilgic, Mustafa, Mihalkova, Lilyana, and Getoor, Lise. Active learning for networked data. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 79–86, 2010.

Budnik, Mateusz, Poignant, Johann, Besacier, Laurent, Quénot, Georges, et al. Automatic propagation of manual annotations for multimodal person identification. In *Proceeding of the 12th International Workshop on Content-Based Multimedia Indexing*, 2014.

Giraudel, Aude, Carré, Matthieu, Mapelli, Valérie, Kahn, Juliette, Galibert, Olivier, and Quintard, Ludovic. The repere corpus: a multimodal corpus for person recognition. In *LREC*, pp. 1102–1107, 2012.

Iwama, Kazuo and Miyazaki, Shuichi. A survey of the stable marriage problem and its variants. In *Informatics*

Education and Research for Knowledge-Circulating Society, 2008. ICKS 2008. International Conference on, pp. 131–136. IEEE, 2008.

Joshi, Hemant, Bayrak, Coskun, and Xu, Xiaowei. Ualr at trec: Blog track. In *TREC*, 2006.

Krumpelman, Chase and Ghosh, Joydeep. Matching and visualization of multiple overlapping clusterings of microarray data. In *Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB'07. IEEE Symposium on*, pp. 121–126. IET, 2007.

Meilă, Marina and Heckerman, David. An experimental comparison of model-based clustering methods. *Machine learning*, 42(1-2):9–29, 2001.

Poignant, Johann, Besacier, Laurent, Quénot, Georges, and Thollard, Franck. From text detection in videos to person identification. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pp. 854–859. IEEE, 2012.

Raghavan, Hema, Madani, Omid, and Jones, Rosie. Active learning with feedback on features and instances. *The Journal of Machine Learning Research*, 7:1655–1686, 2006.

Settles, Burr. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.